Developments in the Built Environment 17 (2024) 100354

Contents lists available at ScienceDirect

Developments in the Built Environment

journal homepage: www.sciencedirect.com/journal/developments-in-the-built-environment

Enhancing occupant experience in defect repair services through text mining-based latent dirichlet allocation metric identification

Jangsoon Kim [a], Wonwoo Shin [b], SangHyeok Han [c], Sungkon Moon [b,*], Jae-Jun Kim [a]

[a] *Department of Architectural Engineering, Hanyang University, Seoul 04763, Republic of Korea*
[b] *Department of Civil Systems Engineering, College of Engineering, Ajou University, Suwon 16499, Republic of Korea*
[c] *Department of Building, Civil and Environmental Engineering, Concordia University, H3G 1M8, Montréal, Québec, Canada*

A R T I C L E I N F O

Keywords:
Evaluation metrics
Occupant experience
Latent dirichlet allocation (LDA)
Repair
Popularity analysis

A B S T R A C T

Given the high demand in South Korea, apartments, constituting 78% of residential projects, have witnessed substantial construction. This growth has coincided with an increase in defects appearing throughout the life cycle of construction projects, frequently becoming a central factor in the escalating disputes between general contractors and occupants. Some defects only surface during occupancy, underscoring the need to address and fulfill occupant requirements during the repairing, given their dissatisfaction can fuel such disputes. Post-occupancy evaluation (POE) offers valuable feedback yet is often subjective, influenced by the perspectives of companies and managers, potentially failing to capture user experiences accurately. The development of POE concerning the construction defect repair service has not been paid attention to, for newly constructed apartments. To overcome these challenges, this paper introduces a Latent Dirichlet Allocation (LDA)-based framework for identifying evaluation metrics, enabling the development of a POE process for construction defect repair services in newly constructed apartments. As a result, “repeated construction”, “rapid response”, “promised fulfillment”, “punctuality”, “defect processing”, and “services and finishing” are identified as key evaluation metrics. Besides enhancing the serviceability of construction defect repair services, the proposed methodology offers the means to continuously monitor occupant experiences and preferences across varying times and regions.

1. Background and introduction

The South Korean economy relies heavily on the construction industry, constituting a substantial portion of its Gross Domestic Product, estimated at 15%–20% (Construction and Economy research Institute of Korea). Apartments stand out as a high demand building type among the various construction projects, accounting for a significant 78% of residential projects due to their popularity in South Korea (Defect Examination Dispute Resolution Committee (DEDR)). The number of defects in newly constructed apartments tends to escalate during the life cycle of construction projects. This paper defines a construction defect as “a deficiency or shortfall in the building’s function, performance, statutory compliance, or user requirements, which may manifest within its structure, materials, services, or other associated facilities” (Watt, 1999).

Given that defects are pervasive not only in apartments but also in other types of facilities, South Korea has established a defect liability system outlined in the Housing Act, the Collective Building Act, the Framework Act on the Construction Industry, and the Multi-Unit Housing Management Act to address and rectify these issues (Korea Ministry of Government Legislationa,b,c,d). Under this system, general contractors, acting as construction companies, offer a warranty period (2–5 years) for newly constructed apartments to repair any defects reported by the occupants during occupancy. Despite these efforts, resolving construction defect disputes remains an ongoing challenge. The Defect Examination Dispute Resolution Committee under the South Korean Ministry of Land has observed a steady rise in such disputes between general contractors and occupants year after year (Defect Examination Dispute Resolution Committee (DEDR)). Furthermore, it is worth noting that numerous researchers concur that construction defect litigation continues to increase worldwide, although comprehensive statistics are not readily available (Brogan et al., 2018; Noble-Allgire, 2009; Grosskopf et al., 2008).

There are primarily two ways to mitigate the number of construction defect disputes: (1) prevent the occurrence of construction defects by identifying and eliminating their root causes in the life cycle of

* Corresponding author.
E-mail address: skmoon@ajou.ac.kr (S. Moon).

https://doi.org/10.1016/j.dibe.2024.100354
Received 17 November 2023; Received in revised form 8 January 2024; Accepted 30 January 2024
Available online 7 February 2024

construction projects, and (2) recognize and satisfy the occupants' requirements during the repair process in the occupancy stage. This paper concentrates explicitly on identifying and fulfilling occupants' needs during repair procedures for construction defects that occur within the warranty period for new apartments. Post-occupancy evaluation (POE) is a method to collect feedback for various parameters, such as energy performance, indoor environment quality (IEQ), occupants' satisfaction, and productivity (Li et al., 2018).

In this respect, Kalantari and Snell (2017) have used POE to examine the effectiveness of design innovations in a mental healthcare facility based on hospital staff perception and feedback acquired via interviews and surveys. This approach is widely adopted to investigate design features tailored to specific occupant groups and the design processes of projects, particularly for public buildings and residential houses (Dikmen and Elias-Ozkan, 2016; Wongbumru and Dewancker, 2016; Grangaard and Ryhl, 2016). The POE methodology measures occupants' satisfaction levels and identifies their requirements; it has been employed to evaluate occupants' comfort, satisfaction, well-being, and health in green buildings (Khair et al., 2015; Martellotta et al., 2016). Furthermore, POE has been utilized by Leder et al. (2016) to explore the factors influencing occupants' satisfaction with the indoor environment within office buildings.

POE techniques can broadly be categorized into subjective methods and physical measurements. The subjective methods involve (1) occupant surveys such as standardized occupant satisfaction, thermal comfort, and customized surveys; (2) interviews and group meetings with occupants or experts; and (3) walkthroughs to identify issues along with photos and/or video recording sessions, checklists, and observation forms. Physical measurements are typically employed for assessing energy and water consumption and IEQ. This includes measurements related to thermal conditions, lighting, acoustics, and other relevant parameters. Subjective methods, such as user surveys and interviews, have been used in previous studies to identify user satisfaction factors of urban infrastructure (e.g., metro and water distribution systems) and improve urban infrastructure maintenance practices for safety and serviceability (Abdul-Rahman et al., 2015; Kang and Lee, 2012; Shafizadeh and Mannering, 2006; Drake and Zechman, 2012a). However, it is noteworthy that there is a gap in the literature regarding conducting POEs for newly constructed apartments to gain a deeper understanding of occupants' experiences of repair services during the warranty period.

In practice, repair services for new apartments are provided in South Korea to repair the defects not detected during construction till a few years after the construction. Upon completing the repairs, a general contractor conducts surveys and interviews with the occupants. These interactions aim to assess the occupants' experiences with the repair services, categorizing their satisfaction levels into "very satisfied", "satisfied", "dissatisfied", and "very dissatisfied". The survey and interview results are stored in a central database in a Korean text format. However, questionnaires used in these surveys and interviews often reflect the construction companies' perspective rather than the occupants' requirements' (Chang et al., 2021; Queirós et al., 2017) since these approaches are subjective, time-consuming, and prone to error. These challenges hinder the serviceability of the repair process and the credibility of construction companies essential to mitigate construction defect disputes and remain competitive in domestic and international construction markets.

To address these limitations, this paper aims to identify evaluation metrics that facilitate an objective assessment of occupants' experiences regarding the defects repair process in newly constructed apartments; it involves applying text-mining techniques to the occupants' comments stored in a Korean text format. To achieve this objective, this paper proposes a quantitative approach consisting of (1) text pre-processing involving cleaning and lemmatization, tokenization, morphological analysis, stopwords removal, and bigrams generation; (2) term frequency (TF) analysis to identify keywords following different levels of occupants' satisfaction and dissatisfaction; and (3) developing a Latent Dirichlet Allocation (LDA) model to extract occupants' topics of interest during repair process of construction defects. This paper uses the occupants' data——obtained from a collaborative construction company in South Korea——in a Korean text format.

2. Literature review

2.1. Post-occupancy evaluation

POE is generally adopted to verify whether the facilities (e.g., buildings, bridges, and roads) perform as intended. To recognize the purposes of POE studies, Li et al. (2018) reviewed large numbers of published papers and the top ten frequently used words, including building(s), performance, occupant(s), evaluate, energy, environment (al), user(s), satisfaction, assess, and indoor. In this respect, POE's aims can be divided into direct and indirect purposes. The direct purposes of POE are to evaluate design, occupants' satisfaction, IEQ (e.g., thermal condition, lighting, and indoor air quality), energy performance (e.g., energy use and energy-saving strategies), and the quality and functionality of facilities (Kalantari and Snell, 2017; Leder et al., 2016; Gupta et al., 2014; Shepley et al., 2012). The indirect purposes of POE involve identifying facilities' technical issues (e.g., overheating risk), improving the POE methods, and evaluating the effectiveness of specific technologies such as mixed-mode air conditioning and natural ventilation (McLeod and Swainson, 2017; Spataru and Gillott, 2011; Thomas, 2017).

Corresponding to the objective of this research, this paper focuses on the direct purposes of POE, especially occupants' satisfaction levels. In a detailed view of direct objectives in POE studies, Ali et al. (2015) explored the effect of the physical environment's comfort on employees' performance in office buildings. Dorsey and Hedge (2017) implemented POE to examine the performance of a LEED platinum-certified building by evaluating the IEQ in relation to the occupants' comfort and satisfaction. Ferri et al. (2015) discussed end-user impressions and experiences in a new intensive care unit built using evidence-based design (e.g., atmosphere and size) through POE. In this respect, POE is executed to understand users' opinions, experiences, and behaviors in various types of facilities such as infrastructure (e.g., railroad stations) and buildings (Chang et al., 2021; Zheng et al., 2010; Brown, 2016). For example, Sanni-Anibire and Hassanain (2016) assessed the quality of student housing facilities related to the performance and control of thermostats, quality of building support services, size of rooms, furniture, and proximity to the cafeteria. Husin et al. (2015) used POE as a safety performance measurement tool to analyze the correlation between occupants' satisfaction and safety performance levels in low-cost housing.

To improve the quality and serviceability of maintenance management for facilities, previous studies focused on applying user experiences obtained from the results of the user surveys (Abdul-Rahman et al., 2015; Kang and Lee, 2012). Drake and Zechman (2012b) sought to learn about the inconveniences (e.g., time and location constraints) of using a water distribution system by analyzing written user complaints. Similarly, Haider et al. (2016) attempted to enhance the reliability of the water supply system by identifying and mitigating technical problems, the leading causes of user dissatisfaction, based on the user complaints in the POE. Chang et al. (2021) examined user experiences and satisfaction levels with infrastructure management for tunnels and bridges under construction by identifying the user-experience factors obtained from civil complaints.

Regarding the failures in building performance, 60% of the most common defects are related to building envelope issues (Brogan et al., 2018). Thus, occupants' satisfaction levels are determined by how efficiently and accurately these defects are repaired (Million et al., 2017). In this respect, previous studies have identified occupant satisfaction factors by exploring the relationship between occupant satisfaction and building performance (Nawawi and Khalil, 2008; Villeneuve and O'Brien, 2020). Based on the results of the studies described above, beyond its general purpose, POE is widely used in the construction domain to evaluate, assess, and investigate occupants' experiences and opinions to improve maintenance strategies and management for better quality, safety, and serviceability.

2.2. Information extraction in text mining

According to advanced technology (e.g., radio frequency tags and 3D laser scanners), the construction industry generates and stores a large set of data such as text documents, numerical data, digital images, and videos for various purposes——heating and air conditioning control for high-quality indoor environment and inspection. Notably, the text documents involving contracts, change orders, design and progress reports, schedules, and accident records are critical; these documents' knowledge and valuable information support the planning, controlling, and decision-making for timely and within-budget completion of construction projects. Therefore, learning to extract useful information from text format-based data efficiently and accurately is essential. The architecture, engineering, and construction (AEC) domain has been focusing on text mining (TM) as one of the solutions since it can summarize knowledge and identify valuable insights from large databases while saving time and cost.

Text mining (TM) is defined as a process where a user analyzes a collection of text data over time, applying a series of techniques (Feldman and Sanger, 2007). In this regard, TM involving text pre-processing and knowledge extraction has been widely used in the AEC domain for safety management, public opinion analysis, compliance checking, building design, method development or improvement, and contract management (Yan et al., 2022). Depending on the objectives of TM projects, the knowledge extraction process involves four categories: (1) text classification, which aims to classify a set of textual documents into predefined categories such as topics and subjects; (2) text clustering, to assign each document in a collection to one or more groups (Feldman and Sanger, 2007); (3) information extraction to extract the required data from the documents to fill in the slots in a pre-defined pattern or spreadsheet (Hassan and Le, 2020); and (4) information retrieval which acts like a search engine retrieving information from a large set of documents based on a query or some keywords. This paper focuses on information extraction from public opinion data, also called user experiences/complaints; it is aligned with this paper's objective of extracting and identifying interesting occupant experience topics in the defects repair process during the warranty period of new apartments.

In the construction industry, public opinions are collected to identify social phenomena; they provide valuable information from social media data, which is enormous, unstructured, and heterogeneous (Yan et al., 2022; Liu and Hu, 2019; Wang et al., 2018) for decision-making, maintenance strategies for green buildings, and off-site construction. In addition, this information can help policymakers understand public attitudes and sentiments effectively. In other words, end-user maintenance feedback contains valuable and worthwhile information to identify the common fault symptoms and problems to improve property management. In this respect, Zhong et al. (2019) developed a convolutional neural network-based approach to explore building quality complaints based on twelve pre-defined parameters, such as leakage, hollowing, cracking, and construction impact.

Topic modeling and sentiment analysis were implemented to extract and understand public issues and user perceptions about infrastructure megaprojects (e.g., metro systems) from social media platforms (Zhang et al., 2021; Zeng et al., 2023). In a similar study, Zeng et al. (Chen et al., 2022) explored public perceptions of construction health and safety by analyzing Instagram posts based on LDA applications. However, these studies did not analyze public opinions in detail to identify the leading causes of users feeling uncomfortable and unsafe while using the infrastructure. To address this challenge, extracting the required information by analyzing sentences and words in user experiences is essential. For instance, Chen et al. (Zhang and El-Gohary, 2016) introduced a framework to understand residential electricity demand using a dynamic energy lifestyles approach. The dynamic energy lifestyles/consumption patterns were obtained by applying LDA to extract latent household energy attributes. To extend the application of information extraction to the construction domain, Zhang and El-Gohary (Moon et al., 2020) developed an automated extraction of requirements from regulatory textual documents based on the application of a rule-based natural language processing approach following syntactic (grammar-related) and semantic (context-related) text features. Moon et al. (Bastani et al., 2019) introduced information extraction frameworks to develop an automated specification reviewing model from bridge inspection reports. Bastani et al. (Jeon et al., 2021) used LDA to extract latent topics from the Consumer Financial Protection Bureau (CFPB) complaint narratives. They explored the associated trends over time to identify insightful knowledge related to the problems experienced by CFPB consumers. On the same lines, Jeon et al. (Hagen et al., 2015) identified metrics to develop a comprehensive Purdue index by analyzing peer-reviewed papers based on the application of LDA.

While previous studies have introduced various TM and POE techniques to enhance building performance and infrastructure maintenance, they still face several limitations. First, current POE practices often rely on subjective questionnaires designed from the perspective of companies or managers, which may not accurately reflect user experiences and satisfaction levels. This gap underscores the need for an objective and rational POE process to effectively identify the underlying issues contributing to occupant satisfaction or dissatisfaction during the repair services within the warranty period. Moreover, occupant priorities regarding construction defect repair services may evolve over time. Consequently, POE evaluation metrics should continuously change to enhance the serviceability and quality of repair services and improve the company's credibility by monitoring occupant experiences. However, no existing study has proposed a framework for systematically monitoring occupant experiences and regularly improving the POE. In response to these challenges and as an initial step towards developing an objective POE tailored to occupants residing in newly constructed apartments, this paper aims to identify evaluation metrics derived from the analysis of occupant experiences presented in a Korean textual format using the LDA algorithm.

3. Proposed methodology

Fig. 1 represents the proposed methodology, which involved data collection, text pre-processing, and identification of evaluation metrics. The initial procedure included collecting occupants' feedback in the text format, which was stored in a central database in a collaborative construction company in South Korea. The collaborative company operates a customer service center (CSC) with the following roles: (1) respond to the occupants' reports regarding defects in their apartment units; (2) coordinate the repairing schedule between the occupants and subcontractors; and (3) collate the occupants' experiences of the repairing services and store their comments into the central database after the

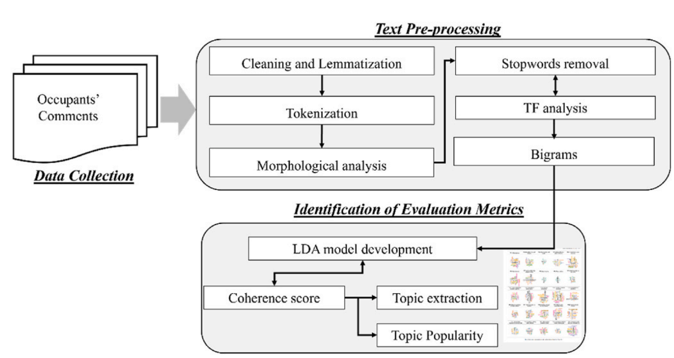

Fig. 1. Research methodology.

repair work ends.

The second step involved text pre-processing, essential in data mining to convert unstructured data into a structured form to extract valuable information efficiently and effectively. This process involves cleaning and lemmatization, tokenization, morpheme analysis, stopword removal, TF analysis, and bigrams generation. At this juncture, it should be noted that the stopword removal and TF analysis are implemented interactively to eliminate irrelevant words (e.g., "painting" and "caulking") related to repairs. The last step, topic modeling, a text mining technique, was implemented to analyze occupant experiences and discover latent topics; this helped develop the evaluation metrics for the POE of repair services for the defects in the newly built apartments. This paper used Python version 3.8.7 to implement the proposed methodology successfully.

3.1. Data collection

Generally, a construction company in Korea provides a warranty period ranging from one to five years to repair the defects reported by the occupants in their newly constructed apartment units, depending on the defects and the project. The procedure for repairing the defects involves (1) the occupants reporting the identified defects to the CSC operated by the general contractor; (2) the CSC requesting the engineers to visit the units to confirm the defects; (3) scheduling repair dates between engineers and occupants after the engineers confirm the defects; and (4) undertaking repair work until the occupants are satisfied with their quality and completeness. To confirm that the occupants are satisfied with the repair work, the CSC receives and stores their feedback in a text format in their database. The feedback contains positive or negative content depending upon the occupants' experiences during the repair services. The raw data had 32,895 sets of feedback (claims to repair the defects) from occupants between October 2018 and September 2021.

As shown in Table 1, the raw data comprised the following information: (1) occupant's feelings and opinions (e.g., "I was disappointed" and "use another brand's product"); (2) the items to be repaired and the problems caused by the defects (e.g., "door closer" and "leakage"); (3) reasons for the occupant's complaints (e.g., "many defects", "no visit and contact", and "not completely removed"); and (4) recurrent repairing services the occupants requested during the repair period (e.g., "in three years I have requested"). In addition, depending on the occupants' experiences during the repair period, they rate the credibility of the company (e.g., "reliability of the company"), which can be significant in improving or damaging the construction company's brand image.

Table 1
Examples of the raw data.

Feedbacks	Raw Data
1	There was no visit and contact even though the visiting date was on June 22nd. Although works to repair the leaking was done previously, leaking was still existed and not fixed yet for a long time. I was very disappointed in the services given by your company since you did not fulfill the promise. The reliability of the company was worst.
2	Although it was paid by me to replace the broken tiles caused by my fault, I hope that you could provide accurate prices estimated in detail such as price per tile and labour cost instead of approximate price. However, I appreciate your service.
3	You did not completely remove the silicone from the glass
4	Why do not you use another brand products since the door was not closed perfectly even though a new door closer, which was the same product used before, was replaced? It was unpleasant that the company did not seem to acknowledge the problems with the initial products.
5	There were too many defects and no repair them quickly. I took off on the weekdays for this repair. Why did not you repair the defects on the weekend? It seems that repairing the floor was done in three years I have requested.

3.2. Text pre-processing

Text pre-processing aims to convert raw text data into meaningful terms for completing the text mining analysis. As illustrated in Fig. 1, the proposed text pre-processing involves six tasks: cleaning and lemmatization, tokenization, morphological analysis, stopwords removal, TF analysis, and bigrams generation. It should be noted that text pre-processing generally involves converting to lowercase in other languages, especially English; however, there are no uppercase and lowercase in Korean. Due to this, this paper does not involve this task. Step 1, cleaning and lemmatization, aims to (1) clean the data of unnecessary noise such as punctuation marks (e.g., "!", "?", ".,", and ",") and numbers, which do not contribute to developing the LDA model; and (2) convert words to their root form (i.e., "cleaning", "cleaned", "cleans" −> "clean"). The spelling check and correct spelling tasks are performed to complete this task successfully.

In addition, the same and similar meaning terms are replaced with a single representative word, leading to reduced feature numbers in the dataset. Tokenization in text preprocessing is critical since it divides a sentence into a set of words called tokens (Naver). Since nouns in Korean commonly provide critical information in each sentence, the research team had to extract meaningful nouns to identify the factors measuring the occupants' interests during/after the repair services (Chang et al., 2021). The morphological analysis aims to tag token morphemes as adjectives, nouns, suffix modifiers, adverbs, or prepositional particles. The nouns are extracted next; however, some Korean adjectives (e.g., "corrective") and verbs (e.g., "repair" and "claim") also contained significant information associated with occupants' interests during the repairing of defects. Therefore, these terms were transformed nouns. For example, "corrective", an adjective, was transformed into "correction", the noun. This paper used the KoNLPy Python package, commonly adopted for successful morphological analysis in Korean.

Based on the results of the morphological analysis, stopword removal was performed to discard common terms with little value for the research objectives. To accelerate the stopword removal, TF analysis, which sorts the most frequent terms efficiently, was employed to identify and eliminate not only the common terms but also the terms involving the statuses of defects (e.g., "leaking" and "broken"), places where there were defects (e.g., "window", "door", "living room", "structure", "entrance", and "bedroom"), and materials (e.g., "drywall", "caulking", "concrete", "light", and "cement"). Finally, the adjacent tokens were combined to create bigrams, primarily used to develop the LDA model (Hagen et al., 2015). Fig. 2 represents the procedures of text pre-processing described above.

3.3. Identification of evaluation metrics using LDA model

Natural language processing (NLP) enables computers to understand natural language as humans do, based on artificial intelligence algorithms. An unsupervised NLP method is widely used in academia and practice to identify patterns from textual data without a time-consuming and tedious labeling process (Wallach, 2006). In this respect, as one of the representative models in the unsupervised NLP method, topic modeling aims to extract latent topics from a large corpus of textual data (Liu et al., 2019). Topic modeling has various techniques, such as LDA, latent and probabilistic latent semantic analyses, based on identifying trending topics and themes, trends of safety accidents, and patterns and trends in the evolution of BIM in the construction domain (Lee et al., 2017; Yalcinkaya and Singh, 2015; Blei, 2012). At this juncture, it should be noted that this paper selected LDA to identify latent topics that became the evaluation metrics for occupants' experience of getting the construction defects repaired during the warranty period of newly built apartments; LDA is less prone to overfitting and demonstrates higher efficiency in dealing with large-scale datasets, also called corpus (Hagen et al., 2015). LDA assigns words in a document to random variables and groups them based on a semantic feature, an iterative probabilistic

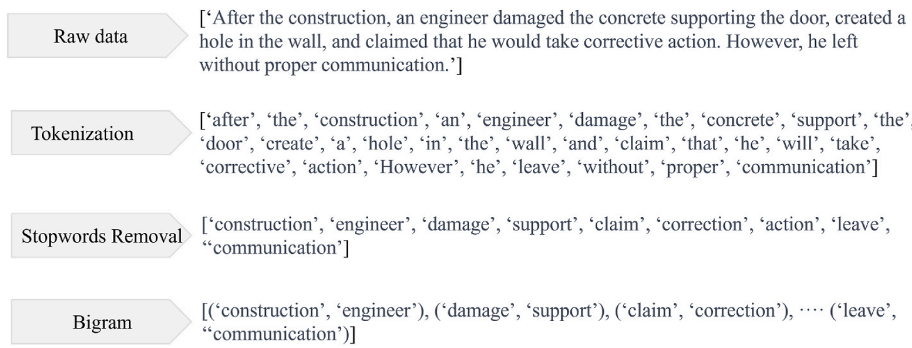

Fig. 2. Procedures of text pre-processing.

process using the Dirichlet distribution (Blei et al., 2003).

Where α, β = dirichlet distribution; δ_k = word distribution in a topic k; θ_d = topic proportion per document d; z = word topic assignment; and w = a specific word.

As shown in Fig. 3, the process flow of the LDA model has three plates that indicate the repetitive processes due to the number of documents (D), the number of words in a single document (N), and the number of topics (K). According to the previous studies (Blei et al., 2003; Stevens et al., 2012), the generation process of the LDA model involves mainly four procedures. First, select a multinomial distribution δ_k for topic k from a Dirichlet distribution with parameter β represented in Eq. (1).

$$\delta_k \sim Dir(\beta) k \in \{1, 2, \ldots\ldots K\} \tag{1}$$

Second, extract a multinomial distribution θ_d for document d from a Dirichlet distribution with parameter α using Eq. (2).

$$\theta_d \sim Dir(\alpha) d \in \{1, 2, \ldots\ldots D\} \tag{2}$$

Third, as shown in Eq. (3), $z_{d,n}$ is derived from the distribution θ_d, and a word, $w_{d,n}$, is computed from the distribution $\delta_{z_{d,n}}$.

$$Z_{d,n} \sim Multi(\theta_d)$$

$$W_{d,n} \sim Multi\left(\delta_{Z_{d,n},n}\right) d \in \{1, 2, \ldots\ldots D\}, n \in \{1, 2, \ldots\ldots N\} \tag{3}$$

Last, in the generative process of LDA, θ_d and $\delta_{Z_{d,n},n}$ are the latent variables but $W_{d,n}$ is the observed variable. As shown in Eq. (4), given α and β as hyperparameters, the marginal probability of documents, D, is computed based on calculating the probability of word, p($W_{d,n}$).

$$p(W_{d,n}|\alpha,\beta) = \int p(\theta_d|\alpha)\left(\prod_{n=1}^{N}\sum_{Z_{d,n}} p(Z_{d,n}|\theta_d)p(W_{d,n}|Z_{d,n},\beta)\right)d\theta_d$$

$$p(D|\alpha,\beta) = \prod_{d=1}^{D}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N}\sum_{Z_{d,n}} p(Z_{d,n}|\theta_d)p(W_{d,n}|Z_{d,n},\beta)\right)d\theta_d \tag{4}$$

To develop an LDA model, α, β, K, and θ must be defined by users depending on the contents of the document set. For example, their values should be more than 1 when the user wants more even distributions for several topics. Otherwise, these values should be less than 1, resulting in concentrated distributions over a few topics or words. To measure the statistical reliability of the LDA model (Sun et al., 2014), the coherence value of the resultant topic model is calculated due to the following roles: (1) assessing the quality of a topic by measuring the degree of semantic similarity between words assigned to a topic; and (2) differentiating between topics which possess semantic interpretability and are artifacts of statistical inference. Therefore, the coherence value is used as a threshold to determine the number of topics based on the error-trial method. The range of the coherence value is between 0 and 1. In this respect, a higher coherence value of a topic model means the topic involves a group of similar words.

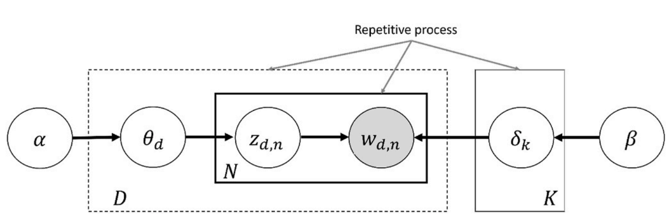

Fig. 3. Process flow of LDA model.

As a result, the number of topics obtained in LDA represents the essential factors that the occupants are interested in to determine their satisfaction levels during the repair services for the remediation of the defects. That is, the topics in this paper will be used as evaluation metrics to assess the repair services objectively so that their quality can be improved continuously by mitigating the causes of occupants' dissatisfaction and enhancing occupant satisfaction. However, the number of topics (i.e., five) cannot be used naturally for the POE of repairing service since it should be simple, easy, and intuitive for optimum occupants' responsiveness. In addition, the factors/topics of occupants' interest can be modified by time periods (in this paper, one month is the time unit). In this respect, this paper implemented the popularity analysis topic, which represents the frequency of the topic over time normalized by the proportions of the topics. Since each occupant feedback was assigned to a mixture of topics with different ratios in the LDA model, the balance must be considered in the topic popularity (P) calculated by Eq. (5).

$$P_{i,t} = \frac{\sum_{t=1}^{n_t}\theta_{i,d_t}}{n_t} i = \{1, 2, \ldots\}, t = \{1, 2, \ldots\}, d = \{1, 2, \ldots\}, \tag{5}$$

where i = topic index; t = time index; d = occupant feedback index; n = total sets of occupant feedback; θ_{i,d_t} = the proportion of topic ID i, assigned to the occupant feedback index d_t.

4. Results and discussion

The initial dataset was sourced from a collaborative construction company. It encompassed a total of 13,895 data points, distributed as follows: 5111 in Seoul, 3351 in Gyeonggi province, 2429 in the central region, 2552 in the southern region, and 452 in Kangwon province within South Korea. To convert this raw data into a meaningful and well-structured format, text pre-processing was conducted using Python. Initially, a cleaning process was applied, which involved two main actions: (1) removal of incomplete or improperly filled occupants' experiences, resulting in the removal of 6842 claims, and (2) elimination of redundant data, including numbers and punctuation marks. To ensure successful lemmatization, an automatic spelling error correction was performed using py-hanspell (National Institute of Korean Language). Correcting spelling errors was deemed essential for effective lemmatization and tokenization in Korean, given its unique grammatical characteristics. It is worth noting that py-hanspell allows users to check and correct a limited number of words (500 words) based on the standard Korean dictionary published by the National Institute of Korean Language (Modupe et al., 2014). Given the dataset's 85,113 space-separated words, spelling error correction was performed repeatedly.

Subsequently, tokenization was employed to break down sentences into individual words, and morphological analysis was conducted to tag the token morphemes using the KoNLPy Python package, specifically the OKT module. Based on tokenization and morphological analysis results, words with similar or identical meanings were consolidated into single representative terms. For instance, as depicted in Table 2, occupants' feedback frequently employed terms like "quickly", "promptly", "fast", "immediately", and "rapid", all of which were substituted with "rapid". A total of 33 synonym cases were identified. It is essential to highlight that nouns were chosen as representative words because, in Korean, nouns typically convey crucial information, such as users' intentions and objectives within each sentence. Significant nouns were extracted from the dataset to align with the research objectives, which involved identifying occupants' areas of interest during and after construction defect repair services. As a result, the dataset ultimately comprised 69,322 words, separated by spaces.

Although prior data cleaning efforts resolved specific issues, the dataset still contained stopwords—words irrelevant to the research objectives. To address this, we employed TF analysis to identify and eliminate a particular set of stopwords from the dataset. These stopwords were determined based on the following criteria: (1) locations and statuses of defects; (2) names of companies; (3) materials; (4) private information of residents (e.g., phone numbers, names, and unit numbers); and (5) emotional behaviors. To elaborate, as presented in Table 2, words like "shout" and "disappointment", representing occupants' emotions, "kitchen" and "window", denoting defect locations, and "brand", signifying company names, were detected and subsequently eliminated. During the TF analysis for identifying stopwords, the term "defect" exhibited the highest frequency. It was designated as a stopword due to its frequent use when occupants reported defects in their units. However, the term "repair" was retained in the dataset because it contained valuable information for understanding occupants' interests efficiently. We created a word cloud to gain insight into occupants' interest in construction defect repair services, as depicted in Fig. 4. The examples of words in the word cloud represented the top ten keywords, which included terms such as "processing", "repair", "time", "request", "response", "appreciation", "construction", "fulfillment", "quality", and "receipt".

The dataset comprised 40,323 words separated by spaces and covered 6234 requisition cases. The research team generated a time frame of requisition cases from October 2017 to September 2021 to gain deeper insights into these data points. Within this timeframe, it was observed that 5165 defect claims occurred between November 2018 and December 2019, with an additional 673 defect cases reported between June 2021 and September 2021. For the remaining periods, there were fewer than 15 defect claims. Regarding the types of defects, the most common were related to doors, windows, general furniture, tiling, heating, ventilation, air conditioning, wallpaper, flooring, and electricity. These defect types accounted for 4751 out of the 6234 total defect claims. This information is illustrated in Fig. 5.

Table 2
Examples of synonyms and stopwords.

Category	word	Representative word
Synonym (33)	"replacement", "recovery", "fix", "repair"	"repair"
	"capacity", "experience", "ability", "skill"	"professionalism"
	"quickly", "promptly", "fast", "immediately", "rapid"	"rapid"
Stopword (436)	"shout", "disappointment", "bathroom", "kitchen", "window", "brand"	

After conducting text pre-processing, the initial step of the LDA analysis involved determining three crucial parameters: α, β, and K. In general, α and β are often referred to as smoothing parameters; they play a significant role in shaping the distribution of topics and words within the dataset (Buenaño-fernandez et al., 2020). When these parameters exceed a value of 1, they tend to yield a more uniform distribution, whereas values below one result in a more concentrated distribution of topics or words. Based on the findings of prior studies, it has been established that setting α and β to 0.1 consistently produces superior results in terms of generating semantically meaningful topics (Jeon et al., 2021; Hagen et al., 2015; Blei et al., 2003). Consequently, this paper adopted the default values of $\alpha = 0.1$ and $\beta = 0.1$.

Furthermore, the number of topics, denoted by K, was utilized as a key input parameter in the LDA model. The optimal value for K was obtained by calculating the coherence value, which assesses the statistical reliability of the resulting LDA model. This coherence value ranges from 0 to 1 and indicates how well a group of closely related words can characterize a topic. A higher coherence value signifies that the LDA model effectively represents the topic with a cluster of coherent words. An extensive range from 1 to 50 was explored to identify the optimal K value. As an initial effort to develop the LDA model, the coherence value was 0.753 when the number of topics was 12. As represented in Table 3, the authors concluded that these topics contained irrelevant words (e.g., common, several, next, name, photo, noise, and forbiddance) or presented repeated keywords (e.g., repair, defect, promised, treatment) in the topics. In this respect, some semantic similarity of the words are grouped by one representative word. For example, "yet", "wait", and "late" are replaced by "delay". The "connection", "installation", and "construction" are considered as "repair". Some irrelevant words are removed. It should be noted that the number of removal stopwords and synonyms represented in Fig. 2 is reported based on result of an initial LDA model development. In addition, as shown in Fig. 1, the LDA model is developed repetitively in accordance with the number of topics resulted by the coherence value.

Based on the replacement of words and elimination of the irrelevant words, remarkably, the highest coherence value of 0.638 was achieved when the number of topics was set to 6. This trend indicates that coherence values decrease as K increases. Consequently, this paper established $K = 6$ as the default value for the LDA model, as it best aligns with the underlying structure of the data. Fig. 6 illustrates the coherence values for different numbers of topics.

The LDA analysis was conducted utilizing the determined parameters mentioned earlier. The results of the LDA analysis, encompassing topics, words, and their proportions (θ), are presented in Table 3. The word proportions signify the frequency of words associated with each topic. To illustrate, consider topic 1, which includes words such as "construction", "repetition", "customer", "inconvenience", "perfection", "response", "repair", and "incomplete", along with their respective proportions enclosed in parentheses. In this context, "construction" holds the highest frequency within topic 1, with a proportion of 0.184. Following closely, "repetition" has the next highest frequency at 0.158. The word "incomplete" exhibits the lowest frequency within this topic. At this juncture, it should be noted that the proportions of the words were also used to label the topics.

To understand and characterize the contents of the topics, it is preferable to label them rather than presenting a combination of words for each topic. Corresponding to the objective of this paper, the identification of evaluation metrics for construction defect repair service labeling the topics is crucial. However, topic labeling, which discovers names of topics in this paper, is generally determined by human interpretation and judgement since LDA is an unsupervised learning process (Hagen et al., 2015; Yalcinkaya and Singh, 2015). That is, it involves a manual process in which the contribution of an expert in the subject is a fundamental element (Nanda et al., 2021). In this respect, after extracting the topics using LDA, the research team had a few meetings

	Word	TF
1	Processing	2774
2	Repair	2555
3	Time	2092
4	Request	1705
5	Response	1336
6	Appreciation	1160
7	Construction	782
8	Fulfillment	777
9	Quality	720
10	Receipt	625

Fig. 4. Word cloud with top 10 keywords.

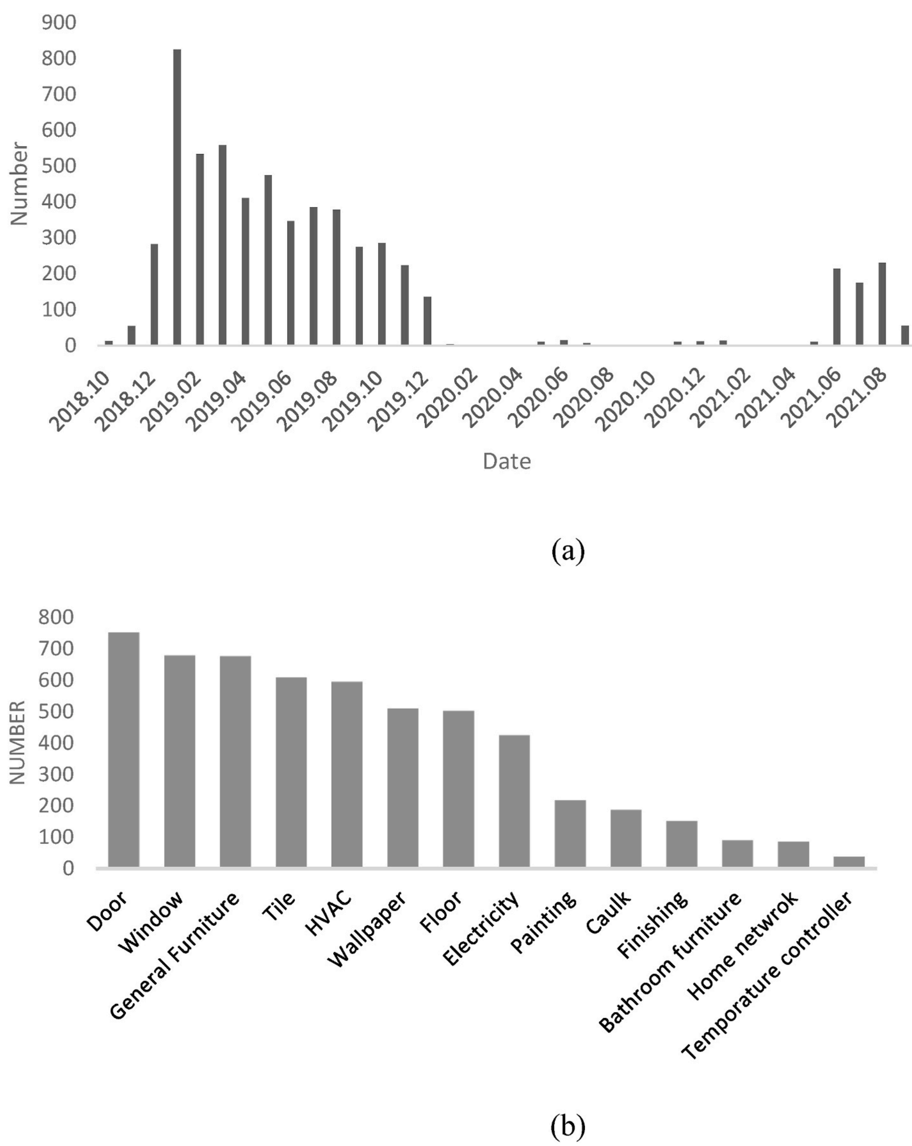

Fig. 5. (a) TF of requisition cases; (b) Types of defects.

with the industrial partner to define the topic labels. Accordingly, as recorded in the third column in Table 4, the following topic labels were determined: repeated construction for topic 1, defect processing for topic 2, services and finishing for topic 3, rapid response for topic 4, promised fulfillment for topic 5, and punctuality for topic 6. At this junction, it should be noted that the industrial partner has constructed various types of buildings in not only South Korea but also the world such as Russia, China, Singapore, and Indonesia. As one of the biggest general contractors, they have mainly constructed high-rise apartments (approximately 12,988 units in past five years) at all regions of South Korea.

To validate these labels, the research team randomly selected occupants' feedback and classified it into one or more relevant labels. This approach was designed to enhance the efficiency and accuracy of the labels to address feedback. To illustrate this validation process effectively, this paper utilized examples of occupants' feedback, as presented

Table 3
Examples of words associated with 12 topics.

Topic	Word
1	Defect, unkindness, completion, continuity, forbiddance, onsite, several, reason
2	Request, average, fulfillment, promised, evaluation, addition, common, reason
3	Time, customer, confirmation, fulfillment, solution, forget, delay, cleanliness
4	Response, defect, next, unkindness, diagnosis, connection, claim, difficulty
5	Kindness, satisfaction, insufficiency, careless, complaint, installation, consumption
6	Kindness, promised, construction, yet, service, training, complaint, noise
7	Visit, promised, wait, rapid, finishing, problem, displeasure, name
8	Company, opinion, rapid, service, displeasure, late, investigation, careless
9	Repetition, defect, repair, solution, treatment, after-service, incompletion, process
10	Treatment, accuracy, late, progress, situation, responsibility, completion, onsite
11	Receipt, quality, complaint, repair, response, wait, photo, smell
12	Appreciation, engineer, cause, construction, inconvenience, solution, temporary, problem

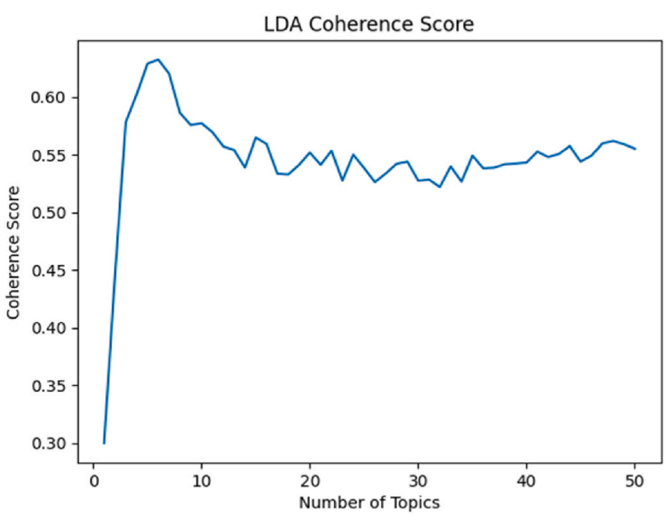

Fig. 6. Coherence values with different number of topics.

Table 4
Extracted topics with associate words and labels.

Topic	Word	Label
1	Construction (0.184), repetition (0.158), customer (0.13), inconvenience (0.11), perfection (0.096), response (0.096), repair (0.048), incompletion (0.045)	Repeated construction
2	Processing (0.362), carefulness (0.134), explanation (0.099), improvement (0.072), repair (0.052), disappointment (0.044), quality (0.036), reliability (0.028)	Defect processing
3	Appreciation (0.176), service (0.102), finishing (0.091), unkindness (0.087), delay (0.075), accuracy (0.072), diligence (0.072), cleanliness (0.066)	Service and finishing
4	Response (0.193), displeasure (0.125), rapid (0.113), satisfaction (0.109), after-service (0.109), quality (0.066), repair (0.06), receipt (0.04)	Rapid response
5	Request (0.142), promised (0.139), engineer (0.123), kindness (0.115), receipt (0.11), problem (0.089), repair (0.043), displeasure (0.038), fulfillment (0.036)	Promised fulfillment
6	Time (0.307), company (0.216), visit (0.184), cleaning (0.088), repair (0.04), quality (0.033), customer (0.022), fulfillment (0.013)	Punctuality

in Table 1. Feedback 1, for instance, can be effectively addressed using the following labels: (1) "punctuality" supported by the statement, "There was no visit and contact even though the visiting date was 22 June."; (2) "repeated construction", exemplified by the statement, "Although the work to repair the leak was done earlier, leaking is still there and not fixed yet for a long time."; and (3) "promised fulfillment" explained by the statement, "I was very disappointed by the services of your company since you did not fulfill the promise.

The reliability of the company is very poor." Feedback 2, in contrast, is best categorized under "defect processing", as indicated by the statement, "I wish you could provide accurate prices estimated in detail, such as price per tile and labor cost, instead of an approximate price." Feedback 3 pertains to "service and finishing", which discusses finishing after repairing the defect. Feedback 4 is closely related to "repeated construction" since it addresses the recurrence of the same defect on a door due to product quality. Lastly, Feedback 5 can be explained by the labels "punctuality", based on the statement, "I took time off on weekdays for this repair but received no visit and contact", and "rapid response" and "defect processing", as supported by the statement, "There were too many defects, and they were not repaired quickly." As a result of this validation process, it is evident that the identified labels effectively addressed the concerns raised in the occupants' feedback.

While we identified six topics using LDA, it is important to note that these topics may not naturally serve as evaluation metrics. This is because occupants' interests in the construction defect repair process can change over time. In other words, a company may not need to prioritize improvements in defect repair services related to a topic trending downwards over time. To address this issue, this paper assessed the popularity of each topic between October 2018 and September 2021. Fig. 7 illustrates the temporal trends of these six topics. These temporal patterns are commonly recognized as indicators of the popularity of their respective topics, which trends of increase, decrease, or variability can categorize. Specifically, the topics "rapid response", "promised fulfillment", and "punctuality" exhibited an increasing trend. "Defect processing" and "services and finishing" garnered attention during specific periods, such as between March 2020 and December 2020, when occupants showed notable interest in these aspects. On the other hand, "repeated construction" maintained consistent attention. It is worth highlighting that Fig. 5(a) presents the period between January 2020 and May 2021, during which the number of defect reports filed by occupants remained below ten. Due to this reason, the trends of the six topics are not adequately represented during this period.

Based on the aforementioned observations and analyses, we can deduce that the critical factors for evaluating occupant experiences encompass the following aspects: (1) repeated repairs leading to occupants' discomfort and inconvenience, as evidenced by words such as "construction", "repetition", "customer", "inconvenience", and "perfection"; (2) defect processing related to the effectiveness of processing defects and supported by the terms, "processing", "carefulness", "explanation", and "improvement"; (3) services and finishing encompassing representative attitudes in the CSC (e.g., "unkindness", "appreciation", and "service") and quality of defect repair (e.g., "finishing", "accuracy", and "cleanliness"); (4) rapid response when defects were reported and indicated by words such as "response", "rapid", "quality", or "repair"; (5) promised fulfillment related to the company's ability to fulfill promises and is associated with words involving "request", "promised", "engineer", "problem" and "fulfillment"; and (6) punctuality indicating the adherence to visiting dates and times, as well as the ability to promptly communicate schedule changes to occupants.

Analyzing the popularity of these factors provides valuable insights into occupant interests and highlights the strengths and weaknesses of the construction defect repair process. Notably, "rapid response", "promised fulfillment", and "punctuality" garnered increased interest over time, suggesting that the company may need to enhance its effectiveness in addressing these aspects of the construction defect repair process. The topic "repeated construction" consistently drew attention,

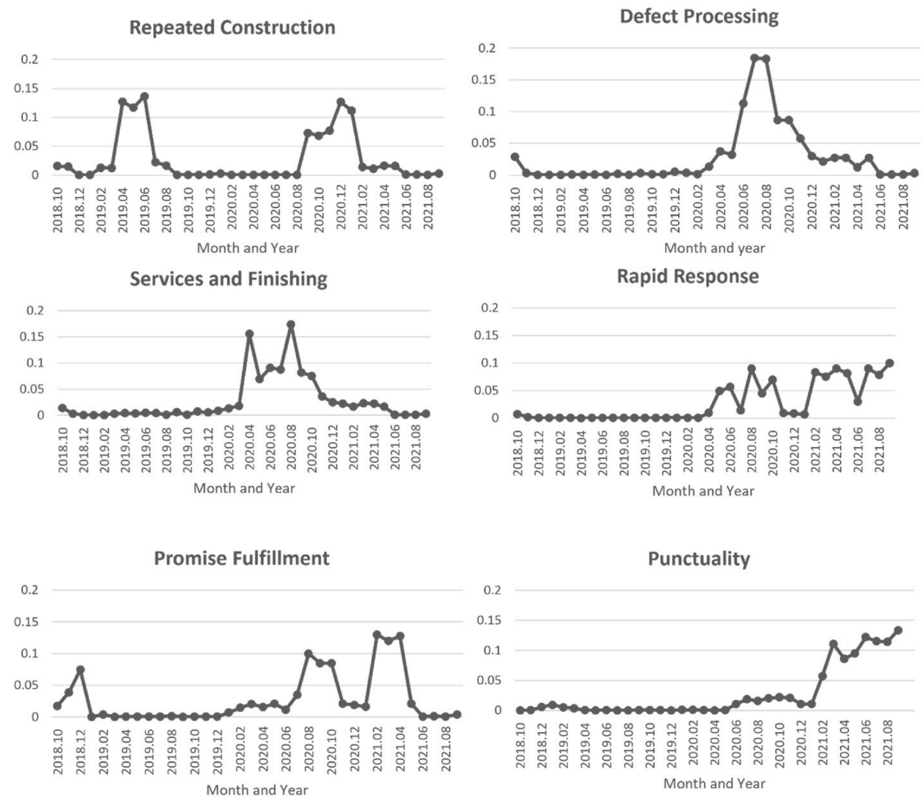

Fig. 7. Result of topic popularity analysis.

implying that the company has not efficiently addressed this issue, as evidenced by a surge in claims between April 2019 and July 2019. Conversely, "defect processing" and "service and finishing" seemed to be on a decreasing trend, indicating that the company has been effectively meeting occupants' requirements in these areas. As a result, analyzing the popularity of these factors not only highlights occupants' interests in the context of construction defect repair services but also reveals areas where the company can improve and excel in its processes.

As future works, this paper involves a few limitations. First, the evaluation metrics identified by this paper is required additional validation with a more extensive corpus of text-based information related to occupant experience in defect repair service, newly built apartments. This is because the trends in topics identified through LDA are not adequately represented in the current dataset, primarily due to its temporal distribution imbalance. For instance, concepts like "defect processing" and "service and finishing" appeared to exhibit a declining trend. Still, their interpretation remains ambiguous due to the limited number of defect claims documented within the specific timeframes (e.g., from January 2020 to May 2021). Furthermore, enhancing the dataset collection process to ensure comprehensive coverage across various regions and to capture recent defect claims is crucial. This is essential because the interests and concerns of occupants may vary over time and across different geographical areas. Consequently, tailoring the construction defect repair process to align with these varying interests becomes an essential consideration in different regions.

Second, previous studies (Hagen et al., 2015; Yalcinkaya and Singh, 2015; Nanda et al., 2021) have reported that topic labeling, which discovers the names of topics, is completed by human interpretation and judgment through the expert experiences and knowledges. Corresponding to this process, this paper determines the names of topics with industrial partner and research team based on the words assigned into each topic. However, this manual approach may involve subjective perception and opinions. To address this limitation, as a future work, qualitative analysis is required to ensure whether or not the topic themes defined by expert and/or researchers are relatively coherent within the words assigned to the topics or in the dataset (Dahal et al., 2019).

Third, as an initial process to develop the POE for construction defect repair services in newly constructed apartments, this paper aims to identify factors as the evaluation metrics, which are main interests of the occupants in the construction defect repair service process, using LDA analysis based on the text-based occupant experiences. Although the popularity analysis represents the trends of occupants' interests over time, this paper does not provide the priorities among the identified factors clearly which should be attention by the company to improve the defect repair service since the dataset used in this paper involves all opinions with satisfaction and dissatisfaction. To address this limitation, sentimental analysis can be used to define the emotional state or opinion of occupants' experiences at each factor into positive (happy) and negative (unhappy) scores [71, 72]. Then, ranking the identified factors is implemented by aggregating these scores to determine which factor (e.g., the highest negative score) should be priority to be improved for better serviceability and quality of the defect repair service.

Lastly, while this study employs the LDA algorithm to discern the factors necessary for developing POEs, it is worth noting that a wide array of text mining techniques, including semantic network analysis and deep learning, can be harnessed to extract these factors from substantial datasets automatically. Subsequently, a comparative analysis of these techniques can be conducted to assess their efficacy in identifying evaluation metrics related to occupants' experiences in construction defect repair processes in newly constructed apartments or other facilities.

5. Summary and conclusion

Given the high demand in South Korea, apartments, constituting 78% of residential projects, have witnessed substantial construction. This trend has coincided with an increase in defects within these newly constructed apartments throughout the life cycle of construction projects. Notably, such defects often serve as a primary catalyst for the persistent rise in construction disputes between general contractors and occupants over the years. While previous studies have identified and eliminated the root causes of construction defects, some are not identified until the occupancy phase. Hence, it becomes imperative to identify and meet the requirements of occupants during the repair process of construction defects at the occupancy stage, recognizing that occupant dissatisfaction can significantly contribute to construction disputes. POE is a valuable tool for collecting feedback for diverse purposes, including assessing energy performance, IEQ, occupants' satisfaction, and productivity. However, POE often takes a subjective form, such as questionnaires, designed from the viewpoint of companies or managers, which may not accurately reflect user experiences and satisfaction. In addition, no study has been paid attention to developing objective POE concerning the construction defect repair service at newly constructed apartments.

To address these combined limitations, this paper endeavored to identify evaluation metrics utilizing the LDA algorithm to reflect occupant experiences concerning the construction defect repair service in newly constructed apartments based on the Korean textual format. The results highlight key evaluation metrics, including "repeated construction", "rapid response", "promised fulfillment", "punctuality", "defect processing", and "services and finishing." Furthermore, a popularity analysis underscored the urgent need for improvement in "rapid response", "promised fulfillment", and "punctuality" within construction defect repair services, as these areas exhibit a concerning upward trend which means that the company does not correspond to the occupants' interests efficiently and effectively. In terms of the proposed approach's effectiveness, it can contribute to (1) continuously monitoring occupant experiences and preferences regarding construction defect repair services, accommodating changes over time and across different regions, and (2) serving as a foundation for ongoing enhancements in the serviceability and quality of construction defect repair services.

CRediT authorship contribution statement

Jangsoon Kim: Writing – original draft, Resources, Formal analysis, Data curation. **Wonwoo Shin:** Writing – original draft, Validation, Methodology, Investigation, Formal analysis. **SangHyeok Han:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Sungkon Moon:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Jae-Jun Kim:** Writing – review & editing, Supervision, Project administration, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (Grant Number: 2022R1F1A1074039).

References

Abdul-Rahman, H.C., Wang, C., Kamaruzzaman, S.N., Mohd-Rahim, F.A., Mohd-Danuri, M.S., Lee, K., 2015. Case study of facility performance and user requirements in the University of Malaya research and development building. J. Perform. Constr. Facil. 29 (5).

Ali, A.S., Chua, S.J.L., Lim, M.E.-L., 2015. The effect of physical environment comfort on employees' performance in office buildings: a case study of three public universities in Malaysia. Struct. Surv. 33 (4/5), 294–308.

Bastani, K., Namavari, H., Shaffer, J., 2019. Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Syst. Appl. 127, 256–271.

Blei, D.M., 2012. Probabilistic topic models. Commun. ACM 55 (4), 77–84.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Brogan, E., McConnell, W., Clevenger, C.M., 2018. Emerging patterns in construction defect litigation: survey of construction cases. J. Leg. Aff. Dispute Resolut. Eng. Constr. 10 (4).

Brown, C., 2016. The power of qualitative data in post-occupancy evaluations of residential high-rise buildings. J. Hous. Built Environ. 31, 605–620.

Buenaño-fernandez, D., Gonzàlez, M., Gil, D., Luàn-mora, S., 2020. Text mining of open-ended questions in self-assessment of University teachers: an LDA topic modeling approach. IEEE Access 8, 35318–35330.

Chang, T., Chi, S., Im, S.-B., 2021. Understanding user experience and satisfaction with urban infrastructure through text mining of civil complaint data. J. Construct. Eng. Manag. 148 (8), 04022061.

Chen, X., Zanocco, C., Flora, J., Rajagopal, R., 2022. Constructing dynamic residential energy lifestyle using Latent Dirichlet Allocation. Appl. Energy 318.

Construction and Economy research Institute of Korea. Construction Industry Image Status and Improvement Plan. <http://www.cerik.re.kr/report/issue/detail/2457> (March 9, 2023) .

Dahal, B., Kumar, S.A., Li, Z., 2019. Topic modeling and sentiment analysis of global climate change tweets. Social Network Anal. Min. 9, 1–20.

Defect Examination Dispute Resolution Committee (DEDR). Defect Inspection Dispute Settlement Casebook. <https://www.adc.go.kr/adms/portal/portalMain.do> (March 9, 2023)..

Dikmen, N., Elias-Ozkan, S.T., 2016. Housing after disaster: a post occupancy evaluation of a reconstruction project. Int. J. Disaster Risk Reduc. 19, 167–178.

Dorsey, J., Hedge, A., 2017. Re-evaluation of a LEED Platinum Building: occupant experiences of health and comfort. Work 57 (1), 31–41.

Drake, K., Zechman, E.M., 2012a. Using consumer complaints to characterize contamination events in a water distribution system. In: Proc., World Environmental and Water Resources Congress. ASCE, Reston, VA, pp. 2247–2252.

Drake, K., Zechman, E.M., 2012b. Using consumer complaints to characterize contamination events in a water distribution system. In: Proceeding World Environmental and Water Resources Congress. ASCE, Reston, VA, pp. 2247–2252.

Feldman, R., Sanger, J., 2007. The TM Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.

Ferri, M., Zygun, D.A., Harrison, A., Stelfox, H.T., 2015. Evidence-based design in an intensive care unit: end-user perceptions. BMC Anesthesiol. 15 (57).

Grangaard, S., Ryhl, C., 2016. Vandhalla – a sport centre and a successful example of first-generation universal design. Stud. Health Technol. Inf. 10 (11).

Grosskopf, K.R., Oppenheim, P., Brennan, T., 2008. Preventing defect claims in hot, humid climates. ASHRAE J. 50 (7).

Gupta, R., Barnfield, L., Hipwood, T., 2014. Impacts of community-led energy retrofitting of owner-occupied dwellings. Build. Res. Inf. 42 (4), 446–461.

Hagen, L., Uzuner, O., Kotfila, C., Harrison, T.M., Lamanna, D., 2015. Understanding citizens' direct policy suggestions to the federal government: a natural language processing and topic modeling approach. In: Proceeding 48th Hawaii International Conference on System Sciences. IEEE, New York, pp. 2134–2143.

Haider, H., Sadiq, R., Tesfamariam, S., 2016. Risk-based framework for improving customer satisfaction through system reliability in small-sized to medium-sized water utilities. J. Manag. Eng. 32 (5).

Hassan, F., Le, T., 2020. Automated requirements identification from construction contract documents using natural language processing. J. Legal Dispute Resol. Eng. Constr. 12 (2).

Husin, H.N., Nawawi, A.H., Ismail, F., Khalil, N., 2015. Correlation analysis of occupants' satisfaction and safety performance level in low cost housing. Procedia – Social Behav. Sci. 168, 238–248.

Jeon, J., Padhye, S., Yoon, S., Cai, H., Hastak, M., 2021. Identification of metrics for the purdue index for construction using latent Dirichlet allocation. J. Manag. Eng. 37 (6).

Kalantari, S., Snell, R., 2017. Post-occupancy evaluation of a mental healthcare facility based on staff perceptions of design innovations. HERD: Health Environ. Res. Des. J. 10 (4).

Kang, K., Lee, K., 2012. Development of a bicycle level of service model from the user's perspective. KSCE J. Civil Eng. 16 (6), 1032–1039.

Khair, N., Ali, H.M., Sipan, I., Juhari, N.H., Daud, S.Z., 2015. Post occupancy evaluation of physical environment in public low-cost housing. J. Teknol. 75, 155–162.

Korea Ministry of Government Legislation. Housing Act. Korean Law Information Center. Available online: https://www.law.go.kr/LSW/eng/engLsSc.do?menuId=2&section=lawNm&query=housing+act&x=0&y=0#liBgcolor38 (accessed on 6 April 2023)..

Korea Ministry of Government Legislation. Act on Ownership and Management of Condominium Buildings. Korean Law Information Center. Available online: https://www.law.go.kr/LSW/eng/engLsSc.do?menuId=2&section=lawNm&query=housing+act&x=0&y=0#liBgcolor2 (accessed on 6 April 2023)..

Korea Ministry of Government Legislation. Framework Act on the Construction Industry. Korean Law Information Center. Available online: https://www.law.go.kr/LSW/eng/engLsSc.do?menuId=2&section=lawNm&query=Framework+Act+on+the+Construction+Industry&x=0&y=0#liBgcolor31 (accessed on 6 April 2023)..

Korea Ministry of Government Legislation. Multi-Family Housing Management Act. Korean Law Information Center. Available online: https://www.law.go.kr/LSW/eng/engLsSc.do?menuId=2&section=lawNm&query=Multi-Family+Housing+Management+Act&x=0&y=0#liBgcolor5 (accessed on 6 April 2023)..
Leder, S.l, Newsham, G.R., Veitch, J.A., Mancini, S., Charles, K.E., 2016. Effects of office environment on employee satisfaction: a new analysis. Build. Res. Inf. 44 (1), 1–17.
Lee, T.Y., Smith, A., Seppi, K., Elmqvist, N., Boyd-Graber, J., Findlater, L., 2017. The human touch: how non-expert users perceive, interpret, and fix topic models. Int. J. Hum. Comput. Stud. 105, 28–42.
Li, P., Froese, T.M., Brager, G., 2018. Post-occupancy evaluation: state-of-the-art analysis and state-of-the-practice review. Build. Environ. 133, 187–202.
Liu, X., Hu, W., 2019. Attention and sentiment of Chinese public toward green buildings based on Sina Weibo. Sustain. Cities Soc. 44, 550–558.
Liu, G., Nzige, J.H., Li, K., 2019. Trending topics and themes in offsite construction research. Construct. Innovat. 19 (3), 343–366.
Martellotta, F., Simone, A., Crociata, S.D., D'Alba, M., 2016. Global comfort and indoor environment quality attributes for workers of a hypermarket in southern Italy. Build. Environ. 95, 355–364.
McLeod, R.S., Swainson, M., 2017. Chronic overheating in low carbon urban developments in a temperate climate. Renew. Sustain. Energy Rev. 74, 201–220.
Million, R.N., Alves, T.D.C.L., Paliari, J.C., 2017. Impacts of residential construction defects on customer satisfaction. Int. J. Build. Pathol. Adapt. 35 (3), 218–232.
Modupe, A., lugbara, O.O., Ojo, S.O., 2014. Filtering of mobile short messaging service communication using latent Dirichlet allocation with social network analysis. In: Transactions on Engineering Technologies. Springer, Dordrecht. Netherlands, pp. 671–686.
Moon, S., Lee, G., Chi, S., Oh, H., 2020. Automated construction specification review with named entity recognition using natural language processing. J. Construct. Eng. Manag. 147 (1).
Nanda, G., Douglas, K.A., Waller, D.R., Merzdorf, H.E., Goldwasser, D., 2021. Analyzing large collections of open-ended feedback from MOOC learners using LDA topic modeling and qualitative analysis. IEEE Trans. Learn. Technol. 14 (2), 146–160.
National Institute of Korean Language. Standard Korean dictionary. https://stdict.korean.go.kr/main/main.do (Accessed on June 15, 2023).
Naver. py-hanspell < https://github.com/ssut/py-hanspell > (Accessed on June 6, 2023).
Nawawi, A., Khalil, N., 2008. Post-occupancy evaluation correlated with building occupants' satisfaction: an approach to performance evaluation of government and public buildings. J. Build. Apprais. 4, 59–69.
Noble-Allgire, A., 2009. Notice and opportunity to repair construction defects: an imperfect response to a perfect storm. Real Prop. Trust Estate Law J. 43 (4).
Queirós, A., Faria, D., Almeda, F., 2017. Strengths and limitations of qualitative and quantitative research methods. Eur. J. Educ. Stud. 3 (9), 1–19.
Sanni-Anibire, M.O., Hassanain, M.A., 2016. Quality assessment of student housing facilities through post-occupancy evaluation. Architect. Eng. Des. Manag. 12 (5), 367–380.
Shafizadeh, K., Mannering, F., 2006. Statistical modeling of user perceptions of infrastructure condition: application to the case of highway roughness. J. Transport. Eng. 132 (2), 133–140.
Shepley, M.M., Rybkowski, Z., Aliber, J., Lange, C., 2012. Ambulatory infusion suite: pre- and post-occupancy evaluation. Build. Res. Inf. 40 (6), 700–712.
Spataru, C., Gillott, M., 2011. The use of intelligent systems for monitoring energy use and occupancy in existing hoes. Sustain. Energy Build. 7, 247–256.
Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D., 2012. Exploring topic coherence over many models and many topics. In: Proceedings of 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, Stroudsburg, PA, pp. 952–961.
Sun, X., Liu, X., Hu, J., Zhu, J., 2014. Empirical studies on the NLP techniques for source code data preprocessing. In: Proceedings of 3rd International Workshop on Evidential Assessment of Software Technologies. Association for Computing Machinery, New York, pp. 32–39.
Thomas, L.E., 2017. Combating overheating: mixed-mode conditioning for workplace comfort. Build. Res. Inf. 45 (1–2), 176–194.
Villeneuve, H., O'Brien, W., 2020. Listen to the guests: text-mining Airbnb reviews to explore indoor environmental quality. Build. Environ. 169, 106555.
Wallach, H.M., 2006. Topic modeling: beyond bag-of-words. In: Proceeding 23rd International Conference on Machine Learning. Association for Computing Machinery, New York, pp. 977–984.
Wang, D., Fan, J., Fu, H., Zhang, B., 2018. Research on Optimization of Big Data Construction Engineering Quality Management Based on RNN-LSTM. Complexity.
Watt, D., 1999. Building Pathology Principles and Practice. Blackwell Science, Oxford, UK.
Wongbumru, T., Dewancker, B., 2016. Post-occupancy evaluation of ser satisfaction: a case study of "old" and "new" public housing schemes in Bangkok. Architect. Eng. Des. Manag. 12 (2), 107–124.
Yalcinkaya, M., Singh, V., 2015. Patterns and trends in building information modeling (BIM) research: a latent semantic analysis. Autom. ConStruct. 59, 68–80.
Yan, H., Ma, M., Wu, Y., Fan, H., Dong, C., 2022. Overview and analysis of the text mining applications in the construction industry. Heliyon 8, 1-16.
Zeng, L., Li, R.Y.M., Yigitcanlar, T., Zeng, H., 2023. Public opinion mining on construction health and safety: latent Dirichlet allocation approach. Buildings 13 (827), 1–21.
Zhang, J., El-Gohary, N.M., 2016. Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking. J. Comput. Civil Eng. 30 (2).
Zhang, Y., Li, D., Li, C., 2021. Public transportation analysis based on social media data. In: Proceeding 19th COTA International Conference of Transportation Professionals. Reston, VA, pp. 1517–1529.
Zheng, M.-C., Chen, M.-S., Li, P.-Y., 2010. Post-occupancy evaluation of information signs and pre-boarding behavior in a historic railroad station. J. Asian Architect. Build Eng. 9 (1), 177–184.
Zhong, B., Xing, X., Love, P., Wang, X., Luo, H., 2019. Convolutional neural network: deep learning-based classification of building quality problems. Adv. Eng. Inf. 40, 46–57.